CSC311 Homework 1

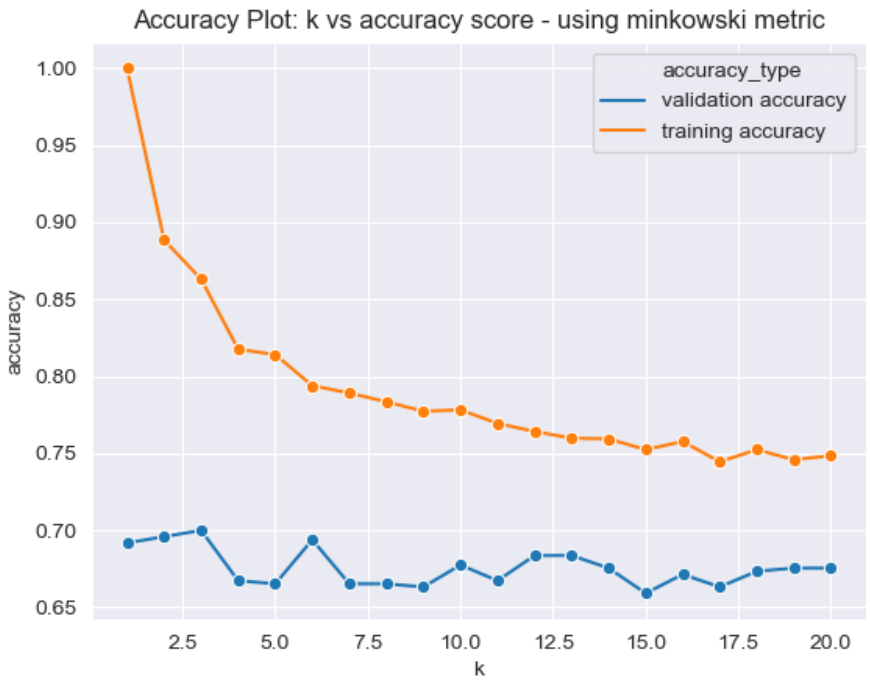
Eric Zhu

30/09/2020

Question 1

Part b

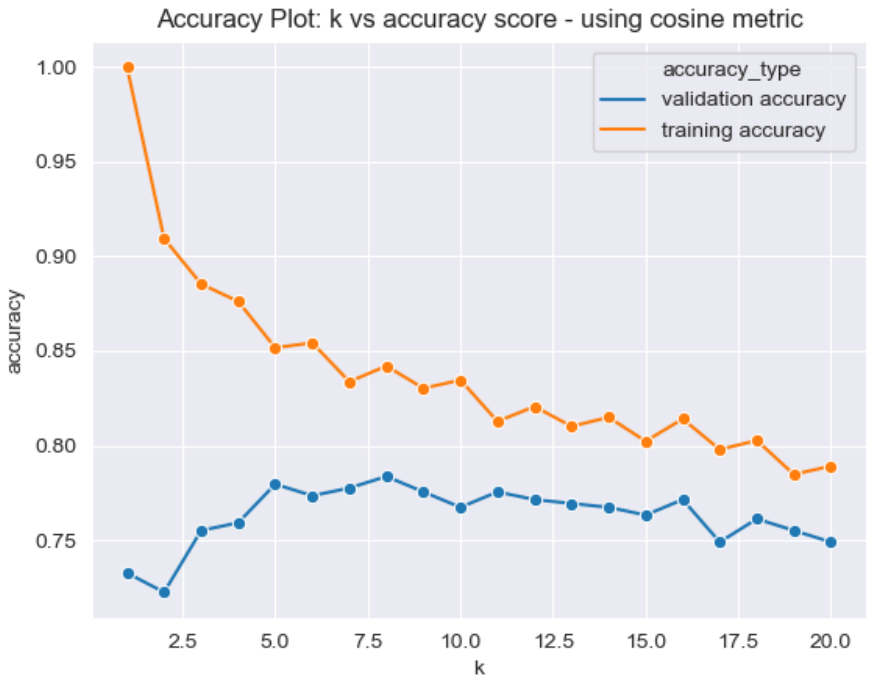
Accuracy plot for the model with the best validation accuracy:



After the 20 observations of models ranging from $k = 1$ to $k = 20$, we see that the model with $k = 3$ produces the model with the best validation accuracy. After calculating the test accuracy on this model, we get a test accuracy of: 0.6979591836734694.

Part c

Accuracy plot for the model with the best validation accuracy:



From switching the metric from "minkowski" to "cosine", we get that our accuracy becomes: 0.7938775510204081, with $k = 8$. We can hypothesize the cause of the increase in accuracy by considering the example dataset: ['cat', 'bulldozer', 'cat cat cat']. Following a similar process to what we used in question 1, i.e, preprocessing the dataset using a CountVectorizer and using a distance-like metric, we'd see that 'cat cat cat' would be 2 units away from 'cat' (3 instances of 'cat' versus just one). However, clearly, 3 instances of cats are more similar to cat rather than bulldozer So we consider the cosine metric that calculates the score based on the angle between the data points. Thus, using this metric, we see that 'cat' and 'cat cat cat' would be very similar since they'd have an angle of 0 between them, i.e, they're scalar multiples of the same vector, while there would be some larger angle in between 'cat cat cat' and 'bulldozer' and similarly between 'cat' and 'bulldozer'. We see that the cosine metric in this example better represents the similarity between the data points.

Question 2

Part a

We wish to derive the update rules for the regularized cost function: $\mathcal{J}_{reg}^{\beta}$, given the regularization function $\mathcal{R}(\mathbf{w}) = \frac{\lambda}{2}\mathbf{w}^T\mathbf{w} = \frac{\lambda}{2}\sum_{j=1}^{D} w_j^2$ and the cost function: $\frac{1}{2N}\sum_{i=1}^{N}(y^{(i)} - t^{(i)})^2$.

First we will derive the update rule for w_j,i.e, $w_j \leftarrow w_j - \alpha\frac{\partial\mathcal{J}}{\partial w_j}$.

Recall that we may split $\mathcal{J}_{reg}^{\beta}$ into $\mathcal{J}^{\beta}(\mathbf{w})$ and $\mathcal{R}(\mathbf{w})$. We will now derive the equation for $\mathcal{J}^{\beta}(\mathbf{w})$:

We start from the linear regression notes:

$$\begin{aligned}\mathcal{J}^{\beta}(\mathbf{w}) &= \mathcal{J}^{\beta}(w_1, w_2, ..., w_j, w_D, b) \\ &= \frac{1}{2N}\sum_{i=1}^{N}(y^{(i)} - t^{(i)})^2\end{aligned}$$

It follows that we can rewrite the above equation in terms of the loss function, i.e:

$$\begin{aligned}\mathcal{J}^{\beta}(\mathbf{w}) &= \frac{1}{2N}\sum_{i=1}^{N}(y^{(i)} - t^{(i)})^2 \\ &= \frac{1}{2N}\sum_{i=1}^{N}\mathcal{L}(y^{(i)}, t^{(i)})\end{aligned}$$

Since $\mathcal{L}(y^{(i)}, t^{(i)})$ is in terms of $\mathbf{w}$, we will calculate the derivative of $\mathcal{L}(y^{(i)}, t^{(i)})$ with respect to w_j, i.e, $\frac{\partial \mathcal{L}(y^{(i)}, t^{(i)})}{\partial w_j}$ and apply the chain rule to find the partial derivative of the unregularized cost function w.r.t w_j. Calculating the partial derivative of the loss function with $y^{(i)} = \sum_{j=1}^{D} w_j x_j^{(i)} + b$, we get:

$$\begin{aligned}\frac{\partial \mathcal{L}(y^{(i)}, t^{(i)})}{\partial w_j} &= \frac{\partial (y^{(i)} - t^{(i)})^2}{\partial w_j} \\ &= 2(y^{(i)} - t^{(i)}) \cdot \frac{\partial y^{(i)}}{\partial w_j}\end{aligned}$$

Thus, we need to calculate $\frac{\partial y^{(i)}}{\partial w_j}$. Note that since $y^{(i)} = \sum_{j=1}^{D} w_j x_j^{(i)} + b$, we can calculate the $\frac{\partial y^{(i)}}{\partial w_j}$ as a linear combination of weights multiplied with observations as follows:

$$\frac{\partial y^{(i)}}{\partial w_j} = \frac{\partial}{\partial w_j}(w_1 x_1^{(i)} + w_2 x_2^{(i)} + w_j x_j^{(i)} + ... + w_D x_D^{(i)} + b) = x_j^{(i)}$$

Now, having calculated $\frac{\partial y^{(i)}}{\partial w_j} = x_j^{(i)}$, we can plug in $x_j^{(i)}$ into $\frac{\partial \mathcal{L}(y^{(i)}, t^{(i)})}{\partial w_j}$ to get:

$$\frac{\partial \mathcal{L}(y^{(i)}, t^{(i)})}{\partial w_j} = 2(y^{(i)} - t^{(i)}) \cdot x_y^{(i)}$$

Thus, we can plug in our result in the above equation for $\frac{\partial \mathcal{L}(y^{(i)}, t^{(i)})}{\partial w_j}$ into $\frac{\partial \mathcal{J}^{\beta}(\mathbf{w})}{\partial w_j}$ to get:

$$\frac{\partial \mathcal{J}^{\beta}(\mathbf{w})}{\partial w_j} = \frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - t^{(i)}) \cdot x_y^{(i)}$$

Having gotten $\frac{\partial \mathcal{J}^{\beta}(\mathbf{w})}{\partial w_j}$, we now derive $\frac{\partial \mathcal{R}(\mathbf{w})}{\partial w_j}$ as follows:

$$\begin{aligned}\frac{\partial \mathcal{R}(\mathbf{w})}{\partial w_j} &= \frac{\partial}{\partial w_j}(\frac{1}{2}\sum_{j=1}^{D}\beta_j w_j^2) \\ &= \frac{\partial}{\partial w_j}(0.5(\beta_1 w_1^2 + \beta_2 w_2^2 + \beta_j w_j^2 + ... + \beta_D w_D^2)) \\ &= \frac{1}{2}\cdot 2 \cdot \beta_j w_j = \beta_j w_j\end{aligned}$$

Thus, we can now add $\frac{\partial \mathcal{R}(\mathbf{w})}{\partial w_j}$ to $\frac{\partial \mathcal{J}^{\beta}(\mathbf{w})}{\partial w_j}$, such that $\frac{\partial \mathcal{J}^{\beta}(\mathbf{w})}{\partial w_j} + \frac{\partial \mathcal{R}(\mathbf{w})}{\partial w_j} = \frac{\partial \mathcal{J}^{\beta}_{reg}(\mathbf{w})}{\partial w_j}$. Expanded, we get:

$$\begin{aligned}\frac{\partial \mathcal{J}^{\beta}_{reg}(\mathbf{w})}{\partial w_j} &= \frac{\partial \mathcal{J}^{\beta}(\mathbf{w})}{\partial w_j} + \frac{\partial \mathcal{R}(\mathbf{w})}{\partial w_j} \\ &= \frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - t^{(i)}) \cdot x_y^{(i)} + \beta_j w_j\end{aligned}$$

We can therefore write:

$$w_j \leftarrow w_j - \alpha(\frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - t^{(i)}) \cdot x_y^{(i)} + \beta_j w_j)$$

Now we repeat a similar procedure to derive $\frac{\partial \mathcal{J}^{\beta}_{reg}(\mathbf{w})}{\partial b}$. In fact, much of the derivation is copy pasted from the derivation of the update rules for w_j.

Recall that the update rule for b: $b \leftarrow b - \alpha\frac{\partial \mathcal{J}}{\partial b}$. Additionally, we may split $\mathcal{J}^{\beta}_{reg}$ into $\mathcal{J}^{\beta}(\mathbf{w})$ and $\mathcal{R}(\mathbf{w})$. We will now derive the equation for $\mathcal{J}^{\beta}(\mathbf{w})$:

We start from the linear regression notes:

$$\begin{aligned}\mathcal{J}^{\beta}(\mathbf{w}) &= \mathcal{J}^{\beta}(w_1, w_2, ..., w_j, w_D, b) \\ &= \frac{1}{2N}\sum_{i=1}^{N}(y^{(i)} - t^{(i)})^2\end{aligned}$$

It follows that we can rewrite the above equation in terms of the loss function, i.e:

$$\begin{aligned}\mathcal{J}^{\beta}(\mathbf{w}) &= \frac{1}{2N}\sum_{i=1}^{N}(y^{(i)} - t^{(i)})^2 \\ &= \frac{1}{2N}\sum_{i=1}^{N}\mathcal{L}(y^{(i)}, t^{(i)})\end{aligned}$$

Since $\mathcal{L}(y^{(i)}, t^{(i)})$ is in terms of $\mathbf{w}$, we will calculate the derivative of $\mathcal{L}(y^{(i)}, t^{(i)})$ with respect to b, i.e, $\frac{\partial \mathcal{L}(y^{(i)}, t^{(i)})}{\partial b}$ and apply the chain rule to find the partial derivative of the unregularized cost function w.r.t b. Calculating the partial derivative of the loss function with $y^{(i)} = \sum_{j=1}^{D} w_j x_j^{(i)} + b$, we get:

$$\frac{\partial \mathcal{L}(y^{(i)}, t^{(i)})}{\partial b} = \frac{\partial (y^{(i)} - t^{(i)})^2}{\partial b}$$
$$= 2(y^{(i)} - t^{(i)}) \cdot \frac{\partial y^{(i)}}{\partial b}$$

Thus, we need to calculate $\frac{\partial y^{(i)}}{\partial b}$. Note that since $y^{(i)} = \sum_{j=1}^{D} w_j x_j^{(i)} + b$, we can calculate the $\frac{\partial y^{(i)}}{\partial b}$ as a linear combination of weights multiplied with observations as follows:

$$\frac{\partial y^{(i)}}{\partial b} = \frac{\partial}{\partial b}(w_1 x_1^{(i)} + w_2 x_2^{(i)} + w_j x_j^{(i)} + ... + w_D x_D^{(i)} + b) = 1$$

Now, having calculated $\frac{\partial y^{(i)}}{\partial b} = 1$, we can plug in 1 into $\frac{\partial \mathcal{L}(y^{(i)}, t^{(i)})}{\partial b}$ to get:

$$\frac{\partial \mathcal{L}(y^{(i)}, t^{(i)})}{\partial b} = 2(y^{(i)} - t^{(i)})$$

Thus, we can plug in our result in the above equation for $\frac{\partial \mathcal{L}(y^{(i)}, t^{(i)})}{\partial b}$ into $\frac{\partial \mathcal{J}^{\beta}(\mathbf{w})}{\partial b}$ to get:

$$\frac{\partial \mathcal{J}^{\beta}(\mathbf{w})}{\partial b} = \frac{1}{N} \sum_{i=1}^{N} (y^{(i)} - t^{(i)})$$

Having gotten $\frac{\partial \mathcal{J}^{\beta}(\mathbf{w})}{\partial b}$, we now derive $\frac{\partial \mathcal{R}(\mathbf{w})}{\partial b}$ as follows:

$$\frac{\partial \mathcal{R}(\mathbf{w})}{\partial b} = \frac{\partial}{\partial b}(\frac{1}{2} \sum_{j=1}^{D} \beta_j w_j^2)$$
$$= \frac{\partial}{\partial b}(0.5(\beta_1 w_1^2 + \beta_2 w_2^2 + \beta_j w_j^2 + ... + \beta_D w_D^2))$$
$$= 0$$

Thus, we can now add $\frac{\partial \mathcal{R}(\mathbf{w})}{\partial b}$ to $\frac{\partial \mathcal{J}^{\beta}(\mathbf{w})}{\partial b}$, such that $\frac{\partial \mathcal{J}^{\beta}(\mathbf{w})}{\partial b} + \frac{\partial \mathcal{R}(\mathbf{w})}{\partial b} = \frac{\partial \mathcal{J}^{\beta}_{reg}(\mathbf{w})}{\partial b}$. Expanded, we get:

$$\frac{\partial \mathcal{J}^{\beta}_{reg}(\mathbf{w})}{\partial b} = \frac{\partial \mathcal{J}^{\beta}(\mathbf{w})}{\partial b} + \frac{\partial \mathcal{R}(\mathbf{w})}{\partial b}$$
$$= \frac{1}{N} \sum_{i=1}^{N} (y^{(i)} - t^{(i)}) + 0 = \frac{1}{N} \sum_{i=1}^{N} (y^{(i)} - t^{(i)})$$

We can therefore write:

$$b \leftarrow b - \alpha(\frac{1}{N} \sum_{i=1}^{N} (y^{(i)} - t^{(i)}))$$

Combined, we see that our two update rules are:

$$w_j \leftarrow w_j - \alpha(\frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - t^{(i)}) \cdot x_y^{(i)} + \beta_j w_j)$$
$$b \leftarrow b - \alpha(\frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - t^{(i)}))$$

Based on this update rule, we see that this form of regularization can be called "weight decay" because based on slide 39 of the week 2 lecture notes, we have that the regularization term here is exactly an L^2 regularized cost, i.e, $\mathcal{J} + \lambda\mathcal{R}$. We know this because be have the extra term $\beta_j w_j$. We can therefore set β to be λ, which therefore decays the weights due to the rearrangement: $(1 - \alpha\beta)w_j - \alpha\frac{\partial\mathcal{J}}{\partial w_j}$.

Part b

We will derive the formulas for $\mathbf{A}_{jj'}$ and $\mathbf{c}_j$ using the simplified linear model $y = \sum_{j=1}^{D} w_j x_j$.

First, recall $\frac{\partial\mathcal{J}_{reg}^{\beta}(\mathbf{w})}{\partial w_j}$:

$$\frac{\partial\mathcal{J}_{reg}^{\beta}(\mathbf{w})}{\partial w_j} = \frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - t^{(i)}) \cdot x_y^{(i)} + \beta_j w_j$$

And then we wish to find $\frac{\partial\mathcal{J}_{reg}^{\beta}(\mathbf{w})}{\partial w_j} = 0$ in the form $\sum_{j'=1}^{D}\mathbf{A}_{jj'}\mathbf{w}_j - \mathbf{c}_j$. First we will rewrite the regularized cost function as:

$$(\frac{1}{N}\sum_{i=1}^{N}x_j^{(i)})(\sum_{j'=1}^{D}w_{j'}x_{j'}^{(i)} - t^{(i)}) + \beta_j w_j$$

It follows that:

$$(\frac{1}{N}\sum_{i=1}^{N}x_j^{(i)})(\sum_{j'=1}^{D}w_{j'}x_{j'}^{(i)} - t^{(i)}) + \beta_j w_j = \frac{1}{N}\sum_{j'=1}^{D}(\sum_{i=1}^{N}x_j^{(i)} \cdot x_{j'}^{(i)}) \cdot w_{j'} - \frac{1}{N}\sum_{i=1}^{N}x_j^{(i)} \cdot t^{(i)} + \beta_j w_j$$
$$= 0$$

Since we do not want $\mathbf{A}_{jj'}$, $\mathbf{c}_j$ to be in terms of w_j, we will define $\delta_{j,j'}$ to be:

$$\delta_{j,j'} = \begin{cases} 1 & j = j' \\ 0 & j \neq j' \end{cases}$$

Next we will expand the terms of $\frac{1}{N}\sum_{j'=1}^{D}(\sum_{i=1}^{N}x_j^{(i)} \cdot x_{j'}^{(i)}) \cdot w_{j'}$ and add $\beta_j w_j$:

$$\frac{1}{N}\sum_{j'=1}^{D}(\sum_{i=1}^{N}x_j^{(i)}\cdot x_{j'}^{(i)})\cdot w_{j'}+\beta_j w_j = \frac{1}{N}((\sum_{i=1}^{N}x_j^{(i)}\cdot x_1^{(i)})\cdot w_1)+\frac{1}{N}((\sum_{i=1}^{N}x_j^{(i)}\cdot x_2^{(i)})\cdot w_2)$$
$$+ ... + \frac{1}{N}((\sum_{i=1}^{N}x_j^{(i)}\cdot x_j^{(i)})\cdot w_j) + ... + \frac{1}{N}((\sum_{i=1}^{N}x_j^{(i)}\cdot x_D^{(i)})\cdot w_D) + \beta_j w_j$$
$$= \frac{1}{N}((\sum_{i=1}^{N}x_j^{(i)}\cdot x_1^{(i)})\cdot w_1)+\frac{1}{N}((\sum_{i=1}^{N}x_j^{(i)}\cdot x_2^{(i)})\cdot w_2)$$
$$+ ... + \frac{1}{N}((\sum_{i=1}^{N}x_j^{(i)}\cdot x_j^{(i)})\cdot w_j) + \beta_j w_j + ... + \frac{1}{N}((\sum_{i=1}^{N}x_j^{(i)}\cdot x_D^{(i)})\cdot w_D)$$

Since we see that since j' takes values from 1 to D, we know that some term in the summation, as we have shown above, will be in terms of j, i.e, $j' = j$. Thus, we can write the above summation in terms of $\delta_{j,j'}$:

$$\frac{1}{N}\sum_{j'=1}^{D}(\sum_{i=1}^{N}x_j^{(i)}\cdot x_{j'}^{(i)})\cdot w_{j'}+\beta_j w_j = \sum_{j'=1}^{D}((\frac{1}{N}\sum_{i=1}^{N}x_j^{(i)}\cdot x_{j'}^{(i)})+\beta_j\delta_{j,j'})\cdot w_{j'}$$

So we can rewrite the partial derivative of the regularized cost function w.r.t w_j as:

$$\frac{1}{N}\sum_{j'=1}^{D}(\sum_{i=1}^{N}x_j^{(i)}\cdot x_{j'}^{(i)})\cdot w_{j'}-\frac{1}{N}\sum_{i=1}^{N}x_j^{(i)}\cdot t^{(i)}+\beta_j w_j = \sum_{j'=1}^{D}((\frac{1}{N}\sum_{i=1}^{N}x_j^{(i)}\cdot x_{j'}^{(i)})+\beta_j\delta_{j,j'})\cdot w_{j'}+\frac{1}{N}\sum_{i=1}^{N}x_j^{(i)}\cdot t^{(i)}$$

Thus, we see that the two parts of the above equation are $\sum_{j'=1}^{D}\mathbf{A}_{jj'}\mathbf{w}_j$ and $\mathbf{c}_j$ respectively, i.e, $\sum_{j'=1}^{D}\mathbf{A}_{jj'}\mathbf{w}_j = \sum_{j'=1}^{D}((\frac{1}{N}\sum_{i=1}^{N}x_j^{(i)}\cdot x_{j'}^{(i)})+\beta_j\delta_{j,j'})\cdot w_{j'}$ and $\mathbf{c} = \frac{1}{N}\sum_{i=1}^{N}x_j^{(i)}\cdot t^{(i)}$. Further we see that $\mathbf{A}_{jj'} = (\frac{1}{N}\sum_{i=1}^{N}x_j^{(i)}\cdot x_{j'}^{(i)})+\beta_j\delta_{j,j'}$, so we conclude **part b**.

Part c

Recall from **part b** that $\mathbf{A}_{jj'} = (\frac{1}{N}\sum_{i=1}^{N}x_j^{(i)}\cdot x_{j'}^{(i)})+\beta_j\delta_{j,j'}$ and that $\mathbf{c}_j = \frac{1}{N}\sum_{i=1}^{N}x_j^{(i)}\cdot t^{(i)}$. Vectorized, we get $\mathbf{A} = \frac{1}{N}\mathbf{X}^T\mathbf{X} + diag(\beta)$ and $\mathbf{c}_j = \frac{1}{N}\mathbf{X}^T\mathbf{t}$, where β is D-dimensional vector of β values.

Note that we are able to vectorize $\mathbf{A}$ as we have done because $\mathbf{X}^T\mathbf{X}$ is P by P matrix ($\mathbf{X}$ is N by P), where P is the dimension of predictors and N is the count of observations/samples. By our linear model $\mathbf{y}$ from the previous parts, we know each predictor has a weight, which implies that $P = D$, i.e, the dimension of predictors is the same as the dimension of weights. Since we know β is a D-dimensional vector, $diag(\beta)$ is a D by D matrix. So we simply add $diag(\beta)$ to $\frac{1}{N}\mathbf{X}^T\mathbf{X}$.

We know our solution to $\mathbf{Aw} - \mathbf{c} = 0$ is $\mathbf{Aw} = \mathbf{c}$ and so it follows that $\mathbf{w} = \mathbf{A}^{-1}\mathbf{c}$. We have:

$$\mathbf{w} = (\frac{1}{N}\mathbf{X}^T\mathbf{X} + diag(\beta))^{-1}(\frac{1}{N}\mathbf{X}^T\mathbf{t})$$

This concludes **part c** and question 2.

Question 3

We wish to derive the 4 following functions: $\mathbf{y}$, $\partial\mathcal{J}/\partial\mathbf{y}$, $\partial\mathcal{J}/\partial\mathbf{w}$, and $\partial\mathcal{J}/\partial b$.

From the course notes, we have in vector form $\mathbf{y} = \mathbf{X}\mathbf{w} + b\mathbf{1}$, where $\mathbf{1}$ is a D-dimensional vector of 1's.

Recall that our loss function is $\mathcal{L}(y,t) = 1 - cos(y-t)$. Component wise, we write our cost function as an average of our loss function over N observations (samples):

$$\mathcal{J} = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(y^{(i)}, t^{(i)})$$

Expanded we get:

$$\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(y^{(i)}, t^{(i)}) = \frac{1}{N}(\mathcal{L}(y^{(1)}, t^{(1)}) + \mathcal{L}(y^{(2)}, t^{(2)}) + ... + \mathcal{L}(y^{(N)}, t^{(N)}))$$

Thus, vectorized we get (note that $cos(\mathbf{A})$ is the element-wise cosine function allowed per Piazza):

$$\mathcal{J} = \frac{1}{N}(\mathbf{1} - cos(\mathbf{X}\mathbf{w} + b\mathbf{1} - \mathbf{t}))\cdot\mathbf{1})$$

We will now find $\partial\mathcal{J}/\partial\mathbf{y}$:

$$\begin{aligned}\frac{\partial J}{\partial \mathbf{y}} &= \frac{\partial}{\partial \mathbf{y}}(\frac{1}{N}(\mathbf{1} - cos(\mathbf{X}\mathbf{w} + b\mathbf{1} - \mathbf{t}))\cdot\mathbf{1})) \\ &= \frac{1}{N}sin(\mathbf{y} - \mathbf{t})\cdot\frac{\partial \mathbf{y}}{\partial \mathbf{y}} \\ &= \frac{1}{N}sin(\mathbf{y} - \mathbf{t})\end{aligned}$$

Next we will find $\partial\mathcal{J}/\partial\mathbf{w}$:

$$\begin{aligned}\frac{\partial J}{\partial \mathbf{w}} &= \frac{\partial}{\partial \mathbf{w}}(\frac{1}{N}(\mathbf{1} - cos(\mathbf{X}\mathbf{w} + b\mathbf{1} - \mathbf{t}))\cdot\mathbf{1})) \\ &= \frac{1}{N}sin(\mathbf{X}\mathbf{w} + b\mathbf{1} - \mathbf{t})\mathbf{X}\end{aligned}$$

Finally, we derive $\partial\mathcal{J}/\partial b$:

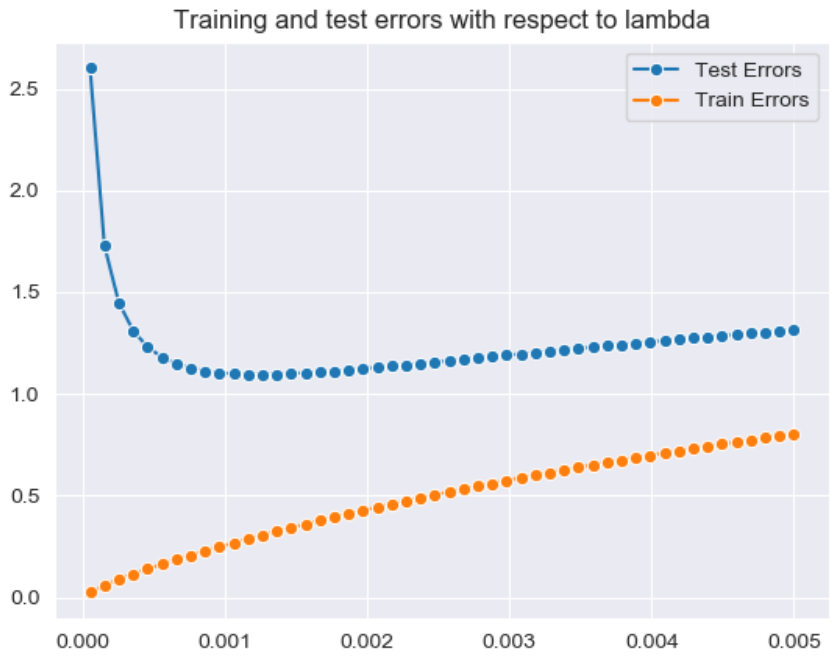
$$\begin{aligned}\frac{\partial J}{\partial b} &= \frac{\partial}{\partial b}(\frac{1}{N}(\mathbf{1} - cos(\mathbf{X}\mathbf{w} + b\mathbf{1} - \mathbf{t}))\cdot\mathbf{1})) \\ &= \frac{1}{N}sin(\mathbf{X}\mathbf{w} + b\mathbf{1} - \mathbf{t})\mathbf{1}\end{aligned}$$

Since we have derived all 4 equations, we conclude question 3.

Question 4

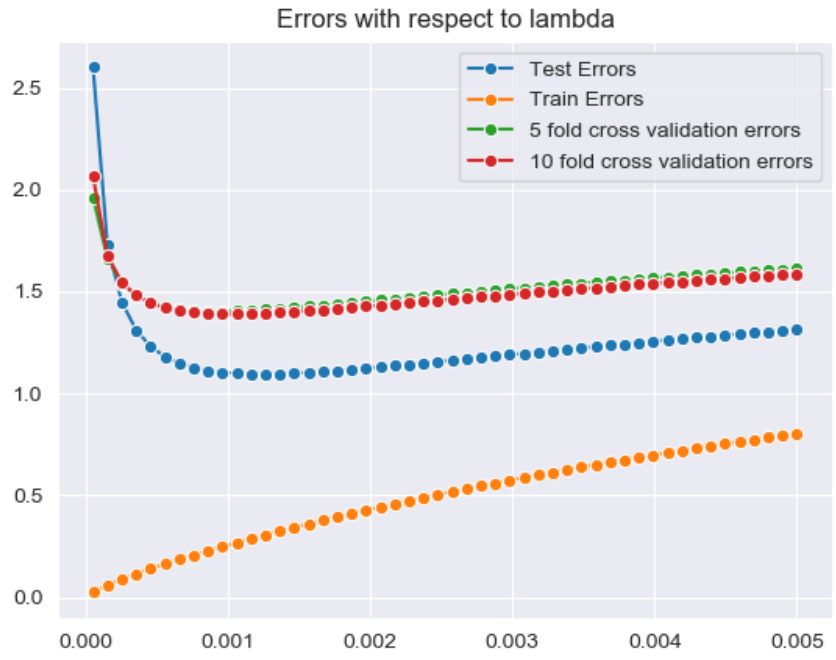
Part c

We will report the training and test errors corresponding to each λ in `lambd_seq` as a plot shown below (suggested by Piazza post 137):



Part d

We plot the training error, test error, 5-fold, and 10-fold cross validation error for each value of `lambd_seq` as follows:



Computationally, we find that for 5-fold cross validation, the minimum CV-error occurs at a value of λ=0.00095918 with a CV-error of 1.40366076, while for 10-fold cross validation, the minimum CV-error occurs at a value of λ=0.00106020 with a CV-error of 1.38968784. We see that 10-fold cross validation produces a smaller minimum error, so we propose that the value of lambda be λ=0.00106020 with a corresponding error of 1.38968784.

Note that we reported the numerical values above to 8 decimal points (and rounded up) for neatness.

We see that the training error is a slowly monotonically increasing curve. Both 5-fold and 10-fold cross validation error curves mirror the shape of the test error curve. We also see that the 10 fold validation curve is an asymptotic lower bound for the 5 fold cross validation error, so it seems that 10-fold performs better than 5-fold cross validation in minimizing errors. Additionally, both error curves generated by k-fold procedures (5,10 folds) are asymptotic upper bounds on the test curve, i.e, both 5-fold and 10-fold cross validation don't seem to perform as well as using just a train and test split as we did in **part c**. Finally, in every error curve except for the training error curve, we see an initial fast decrease in error as λ increases, then the error very slightly increases for increasing λ after λ is sufficiently large.